Thank God That Regressing Y on X is Not the Same as Regressing X on Y: Direct and Indirect Residual Augmentations

Thank God That Regressing Y on X is Not the Same as Regressing X on Y: Direct and Indirect Residual Augmentations

Xiaojin Xu and Xiao-Li Meng*
Department of Statistics, Harvard University
Yaming Yu
Department of Statistics, University of California, Irvine

October 24, 2012

Abstract

What does regressing Y on X versus regressing X on Y have to do with MCMC? It turns out that many strategies for speeding up data-augmentation type algorithms can be understood as fostering independence or "de-correlation" between a regression function and the corresponding residual, thereby reducing or even eliminating dependence among MCMC iterates. There are two general classes of algorithms, those corresponding to regressing parameters on augmented data/auxiliary variables and those that operate the other way around. The interweaving strategy (Yu and Meng, 2011, *JCGS*) provides a general recipe to automatically take advantage of both, and it is the existence of two different types of residuals that makes the interweaving strategy seemingly magical in some cases and promising in general. The concept of residuals—which depends on actual data—also highlights the potential for substantial improvements when data augmentation schemes are allowed to depend on the observed data, a potential that so far has been largely overlooked. At the same time, there is an intriguing phase transition type of phenomenon regarding choosing (partially) residual augmentation schemes, reminding us once more of the prevailing issue of trade-off between robustness and efficiency. This article reports on these latest theoretical investigations (using a class of normal/independence models) and empirical findings (using a posterior sampling for a Probit regression) in the search for effective residual augmentations—and ultimately more MCMC algorithms—that meet the 3-S criterion: *simple, stable, and speedy.*

Keywords: Ancillary-Sufficient Interweaving Strategy (ASIS), Conditional Augmentation, MCMC, Marginal Augmentation, Phase transition, Probit Regression, PX-DA,

1 Residual Augmentations: A Unified Strategy

1.1 Creative Re-Parameterization and Over-Parameterization

Designing algorithms that are *simple, stable, and speedy* is a dream shared by virtually anyone working on Markov chain Monte Carlo (MCMC) or more generally on statistical computing. For data augmentation (Tanner and Wong 1987) and Gibbs sampling (Geman and Geman 1984; Gelfand and Smith 1990) type of algorithms, it is well known that parameterizations can affect substantially both convergence and ease of implementation (e.g., Gelfand, Sahu and Carlin 1995, 1996; van Dyk and Meng, 2010). By

*Corresponding author: *meng@stat.harvard.edu*

using parameterizations creatively, a variety of strategies have been proposed to accelerate convergence while maintaining implementation simplicity. In particular, Papaspiliopoulos, Roberts and Sköld (2003, 2007) study the *centered*, *noncentered*, and *partially noncentered* parameterizations. The idea of partial noncentering is to introduce a family of parameterizations (or data augmentation schemes), and then to seek the optimal parameterization for fastest convergence. This is mathematically equivalent to the conditional augmentation approach (Meng and van Dyk, 1999; van Dyk and Meng, 2001), where the family of data augmentation schemes are indexed by a *working parameter*.

Formally, consider the model $p(\theta|Y_{obs}) \propto p(Y_{obs}|\theta)p(\theta)$ where θ is the parameter of interest and Y_{obs} denotes observed data. A data augmentation (DA) model $p(Y_{obs}, Y_{mis}|\theta)$ is any joint distribution of Y_{mis} (the missing or augmented data) and Y_{obs} given θ such that the marginal $p(Y_{obs}|\theta)$ is preserved. In other words, we can write $p(Y_{obs}, Y_{mis}|\theta) = p(Y_{obs}|\theta)p(Y_{mis}|Y_{obs}, \theta)$. In conditional augmentation, a working parameter c is introduced such that

$$p(Y_{obs}, Y_{mis}|\theta, c) = p(Y_{obs}|\theta)p(Y_{mis}|Y_{obs}, \theta, c).$$

Whereas $p(Y_{obs}, Y_{mis}|\theta, c)$ clearly is a legitimate DA because it preserves the desired margin $p(Y_{obs}|\theta)$, it is a form of over-parameterization because the working parameter c is not identifiable by the observed data Y_{obs}. For conditional augmentation, the value of c is obtained by optimizing a certain criterion, e.g., based on the convergence rate of the closely related EM algorithm (Meng and van Dyk, 1997, 1998, 1999 and van Dyk and Meng, 2001). The resulting algorithm alternates between drawing θ given (Y_{obs}, Y_{mis}) and drawing Y_{mis} given (θ, Y_{obs}), conditioning on the chosen value of c. Finding a good conditional augmentation scheme requires a careful balance between the theoretical speed and ease of implementation, as illustrated in detail by van Dyk and Meng (2001, 2010).

This *conditional augmentation* approach contrasts with the *marginal augmentation* approach (Meng and van Dyk, 1999), which is closely related to parameter-expanded DA (PX-DA; Liu and Wu, 1999). In marginal augmentation, the working parameter c is marginalized out after being assigned a *working prior* $p(c)$. The resulting algorithm is a standard DA—labeled Scheme 2 in van Dyk and Meng (2001)—alternating between drawing Y_{mis} given (θ, c, Y_{obs}) and drawing (θ, c) given (Y_{mis}, Y_{obs}) based on the joint posterior

$$p(Y_{mis}, \theta, c|Y_{obs}) \propto p(Y_{obs}, Y_{mis}|\theta, c)p(\theta)p(c). \tag{1.1}$$

We can also sample from (1.1) by alternating between drawing (Y_{mis}, c) given (θ, Y_{obs}) and drawing (θ, c) given (Y_{mis}, Y_{obs}), as in PX-DA (Liu and Wu, 1999). Obviously this is algorithmically equivalent to the DA sampler that alternates between drawing Y_{mis} given (θ, Y_{obs}) and drawing θ given (Y_{mis}, Y_{obs}), which was labeled Scheme 1 in van Dyk and Meng (2001).

The strategies discussed above all amount to using a single data augmentation scheme in the actual implementation. For conditional augmentation, this is rather obvious by construction. For marginal

augmentation, if the working prior $p(c)$ is proper, then Scheme 1 is the standard DA using

$$\tilde{p}(Y_{mis}|Y_{obs},\theta) = \int p(Y_{mis}|Y_{obs},\theta,c)p(c)\mu(dc) \tag{1.2}$$

as the data augmentation, where μ is the dominating measure for the working prior, typically the Lebesgue measure. However, when $p(c)$ is improper, Scheme 1 is not feasible. In contrast, Scheme 2 still is implementable, just as an improper prior can still lead to a proper posterior. But this does not automatically imply that the algorithm will converge properly. Minimally it should be clear that the resulting joint chain for (θ, c, Y_{mis}) cannot be positive recurrent because its target distribution (1.1) is improper when $p(c)$ is improper. By a result of Hobert (2001a, b), this also automatically implies that the corresponding (major) sub-chain for (θ, c) cannot be positive recurrent either. However, when the improper working prior is the limit of a sequence of proper priors, then under regularity conditions, the *sub-sub-chain* produced by Scheme 2 for θ will converge to the desired target distribution $p(\theta|Y_{obs})$. Intriguingly, when $p(c)$ corresponds to the right Haar measure, this sub-sub-chain actually represents the fastest algorithm among a class of DA algorithms as formulated in Liu and Wu (1999) with their elegant group-theoretic argument.

Even more intriguingly, there is often a simpler way to reach this optimality by using two standard data augmentation schemes (i.e., no improper prior is involved), and the new strategy is demonstrably more powerful and versatile than all known strategies based on a single (limiting) data augmentation, for reasons presented in the following section.

1.2 Alternating versus Interweaving

Suppose $p(Y_{mis}, \theta|Y_{obs})$ and $p(\tilde{Y}_{mis}, \theta|Y_{obs})$ are two augmentation schemes (i.e., both preserving the target posterior $p(\theta|Y_{obs})$). An obvious strategy is to concatenate two iterations, one based on each of the two schemes, that is, by *alternating* between the two algorithms. This may be represented schematically as in Figure 1 where each arrow indicates a sampling step. For example $Y_{mis} \to \theta$ means drawing θ given the current Y_{mis} (and Y_{obs}). Somewhat surprisingly, Yu and Meng (2011) demonstrate that an alternative *interweaving strategy* holds much more promise than the simple alternating scheme. Specifically, the interweaving strategy simply cuts out the θ between Y_{mis} and $\tilde{Y}_{mis}$, and hence it leads to the triangular diagram given in Figure 2. That is, each iteration cycles through the parameter θ and the two sets of augmented data by first drawing Y_{mis} given θ, then $\tilde{Y}_{mis}$ given Y_{mis}, and then θ given $\tilde{Y}_{mis}$. (Henceforth we suppress the conditioning on Y_{obs} when there is no confusion.)

The triangular diagram also reveals a fundamental insight about the power of the interweaving strategy. Similar to the usual DA algorithm, whose convergence rate is the square of the maximal correlation between Y_{mis} and θ in their joint posterior, the interweaving strategy has a convergence rate that is bounded above by the product of three maximal correlations as indicated by the three links in the above diagram. That is, let the geometric convergence rate of DA under Y_{mis} and $\tilde{Y}_{mis}$ be r_1 and r_2,

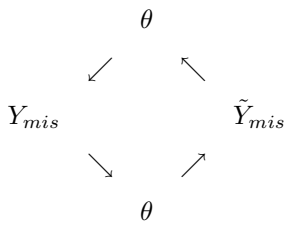

Figure 1: Alternating Scheme

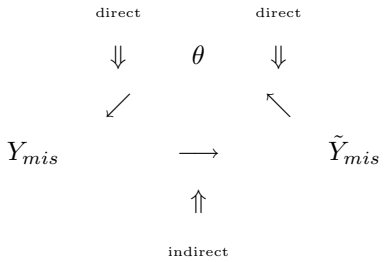

Figure 2: Interweaving Scheme

respectively, and the rate for the interweaving scheme be $r_{1\&2}$. Then Yu and Meng (2011) proved that

$$r_{1\&2} \leq \mathcal{R}(Y_{mis}, \tilde{Y}_{mis})\mathcal{R}(\theta, Y_{mis})\mathcal{R}(\theta, \tilde{Y}_{mis}) = \mathcal{R}_{1,2}\sqrt{r_1 r_2} \tag{1.3}$$

where $\mathcal{R}_{1,2} \equiv \mathcal{R}(Y_{mis}, \tilde{Y}_{mis})$, and

$$\mathcal{R}(X_1, X_2) = \sup_{g,h \in L^2} \text{Corr}\{g(X_1), h(X_2)\}$$

is the maximal correlation between (generic) X_1 and X_2. (Note in our application, the joint distribution is the joint posterior predictive distribution $p(Y_{mis}, \tilde{Y}_{mis}|Y_{obs})$.)

As discussed in Yu and Meng (2011), the key insight here is that we can make $r_{1\&2}$ small (which means a faster algorithm) by making *any one of* $\{\mathcal{R}_{1,2}, r_1, r_2\}$ small. Indeed, it is even possible that $r_1 = r_2 = 1$, that is, neither of the two DAs being interwoven is geometrically convergent, and yet $r_{1\&2} = 0$, that is, the interwoven algorithm will deliver i.i.d draws! See Yu and Meng (2011) for such an example.

In general, achieving i.i.d. draws is obviously too much of a dream, but the interweaving strategy provides us with a new way to combat the common problem of high dependence among consecutive MCMC draws. Specifically, with either alternating or interweaving, we can reduce the dependence between $\theta^{(t)}$ and $\theta^{(t+1)}$—where t indexes the iteration—by reducing either r_1 or r_2 or both. Schematically, this corresponds to "breaking" either or both of the two *direct* links marked in Figure 2; here a direct link is an arrow with θ as one of its two end points. However, the interweaving strategy allows us to achieve the same goal by breaking an *indirect* link, which does not involve θ, and clearly it exists only in Figure 2.

Therefore, given the original augmentation as represented by the arrow from θ to Y_{mis}, we now have two ways to break the cycle. The first is to make $\tilde{Y}_{mis}$ independent of θ and hence to break the $\tilde{Y}_{mis} \to \theta$ link, which is what partially non-centering or conditional augmentation aims to achieve. The second is to make $\tilde{Y}_{mis}$ independent of Y_{mis}, thereby breaking the $Y_{mis} \to \tilde{Y}_{mis}$ link, which is what marginal augmentation and the interweaving strategy try to accomplish. In particular, Yu and Meng (2011) advocate an ancillarity-sufficiency interweaving strategy (ASIS) that takes advantage of the existing competing nature between sufficient augmentation and ancillary augmentation to reduce their *a posteriori* dependence.

1.3 Direct and Indirect Residual Augmentations

Such considerations lead to the idea of residual augmentations (Yu and Meng 2011, Rejoinder), as a way to "break links" by judiciously choosing $\tilde{Y}_{mis}$ for a given (original) DA scheme (Y_{mis}, θ). For the *direct residual augmentation* (DRA), we attempt to break the direct link $\tilde{Y}_{mis} \rightarrow \theta$ by choosing $\tilde{Y}_{mis}$ to be a residual from regressing Y_{mis} on θ. The central idea here is that a residual is constructed to be uncorrelated (though rarely independent) with the regression function, which is θ here. The obvious choice is the usual additive residual from regressing Y_{mis} on θ :

$$\tilde{Y}_{mis} = Y_{mis} - \mathrm{E}[Y_{mis}|\theta, Y_{obs}]. \tag{1.4}$$

A less obvious one is its multiplicative variant:

$$\tilde{Y}_{mis} = \frac{Y_{mis}}{\mathrm{E}[Y_{mis}|\theta, Y_{obs}]} \tag{1.5}$$

in the scalar case. It is straightforward to show that both $\tilde{Y}_{mis}$'s are uncorrelated with θ with respect to the joint posterior distribution $p(Y_{mis}, \theta|Y_{obs})$, as long as the correlation exists. (But note the condition of having correlation does not hold for (1.5) as often as it does for (1.4)).

For the *indirect residual augmentation* (IRA), the aim is to break the indirect link $Y_{mis} \rightarrow \tilde{Y}_{mis}$, and hence we need to regress θ on Y_{mis}. This naturally leads to the counterparts of (1.4) and (1.5) by swapping θ and Y_{mis}, that is,

$$\tilde{Y}_{mis} = \theta - \mathrm{E}[\theta|Y_{mis}, Y_{obs}] \tag{1.6}$$

and

$$\tilde{Y}_{mis} = \frac{\theta}{\mathrm{E}[\theta|Y_{mis}, Y_{obs}]}. \tag{1.7}$$

For all these constructions, the implementation $Y_{mis} \rightarrow \tilde{Y}_{mis}$ is typically straightforward. We accomplish this by first drawing θ from $p(\theta|Y_{mis}, Y_{obs})$, which is a step required by the original DA algorithm based on Y_{mis} alone. We can then compute $\tilde{Y}_{mis}$ as a deterministic function of θ, Y_{mis} and Y_{obs}. This computation typically is straightforward for DRA, because $\mathrm{E}[Y_{mis}|\theta, Y_{obs}]$ is simply the mean function of the full conditional $p(Y_{mis}|\theta, Y_{obs})$ already needed by the original DA algorithm; it can also be carried out by Monte Carlo if necessary. For IRA, this task typically is even simpler, because it calls only for $\mathrm{E}[\theta|Y_{mis}, Y_{obs}]$, the complete-data posterior mean.

Therefore, the simplicity of a residual augmentation algorithm depends critically on how easy it is to implement the $\tilde{Y}_{mis} \rightarrow \theta$ step. To implement it exactly requires us to derive the conditional distribution of θ given $\tilde{Y}_{mis}$ as implied by one of (1.4)-(1.7). This may not be an easy task when the regression function involved (i.e., $\mathrm{E}[Y_{mis}|\theta, Y_{obs}]$ or $\mathrm{E}[\theta|Y_{mis}, Y_{obs}]$) is non-linear. This issue, however, can be dealt with pragmatically by adopting a convenient global or local approximation, with the trade-off of achieving less reduction in auto-correlations for implementation simplicity. Such a pragmatic approach also helps us to compromise appropriately between implementation simplicity and the desire to find

suitable transformations of $g(\theta)$ and $h(Y_{mis})$ such that the low correlation between them is a reasonable indicator of their lack of dependence. Note ideally we would want a joint one-to-one transformation $T(\theta, Y_{mis})$ for better joint normality because under joint normality low linear correlation is the same as low maximal correlation. Unfortunately, this joint transformation typically will destroy the simplicity of the original Gibbs setup that alternates between θ and Y_{mis}.

For the rest of the paper, in Section 2 we first illustrate some theoretical properties of residual augmentations using the simplest normal hierarchical model and its extensions, which include t distributions. In particular, we note an interesting "safe zone" for the choice of augmentation schemes and show how ASIS can be viewed as a "minimax" strategy, always staying within the safe zone regardless of the prior specification and the configuration of observed data. Our pragmatic strategy is illustrated in Section 3 with a probit regression example. We conclude in Section 4 with a host of open problems.

2 Theoretical Illustrations and a Phase Transition Phenomenon

2.1 Illustrating DRA and IRA

A common illustrative example in the DA literature is the one-way random effect model (Liu and Wu, 1999; Yu and Meng, 2011; Hobert and Roman, 2011). Instead of repeating the standard setup, here we adopt a simpler representation capturing its essence that is relevant for our algorithmic investigation. Specifically, suppose θ is the parameter of interest and Y_{mis} is the missing datum or latent variable, and their joint posterior distribution (given Y_{obs}) can be standardized into

$$\left(\begin{array}{c} \theta \\ Y_{mis} \end{array} \right) \Bigg| Y_{obs} \sim N \left[\left(\begin{array}{c} 0 \\ 0 \end{array} \right), \left(\begin{array}{cc} 1 & r \\ r & 1 \end{array} \right) \right]. \tag{2.1}$$

Here r is a known function of Y_{obs} and, without loss of generality, we can assume $0 \leq r < 1$. The standard DA based on Y_{mis} then iterates between sampling θ given Y_{mis} and sampling Y_{mis} given θ (all conditioning on Y_{obs} of course). Clearly this DA has the convergence rate $r_1 = r^2$.

Now consider a conditional augmentation or partially non-centering scheme $\tilde{Y}_{mis} = Y_{mis} - c\theta$, with c being a working parameter to be determined. Clearly

$$\left(\begin{array}{c} \theta \\ \tilde{Y}_{mis} \end{array} \right) \Bigg| Y_{obs} \sim N \left[\left(\begin{array}{c} 0 \\ 0 \end{array} \right), \left(\begin{array}{cc} 1 & r-c \\ r-c & 1+c^2-2rc \end{array} \right) \right]. \tag{2.2}$$

This implies that the DA algorithm using $\tilde{Y}_{mis}$ as the augmentation will have convergence rate $r_2 = (r-c)^2/(1+c^2-2rc)$. Now because of their joint normality, the maximal correlation between Y_{mis} and $\tilde{Y}_{mis}$ is the same as the absolute value of their linear correlation. Therefore $\mathcal{R}_{1,2} = |\text{Corr}(\tilde{Y}_{mis}, Y_{mis})| = |1-cr|/\sqrt{1+c^2-2rc}$. Because the bound in (1.3) is sharp for this normal setting (Yu and Meng, 2011), we see that the rate of convergence from interweaving the DA based on Y_{mis} and the DA based on $\tilde{Y}_{mis}$

is

$$r_{1\&2} = \mathcal{R}_{1,2}\sqrt{r_1 r_2} = \frac{|r(r-c)(1-cr)|}{1+c^2-2rc}. \tag{2.3}$$

We see immediately that when $c = r$ or $c = r^{-1}$, $r_{1\&2} = 0$, and hence the interweaving strategy will produce i.i.d. draws. The $c = r$ case corresponds to DRA because $\mathrm{E}[Y_{mis}|\theta, Y_{obs}] = r\theta$, and hence taking $c = r$ in $\tilde{Y}_{mis} = Y_{mis} - c\theta$ is the same as making $\tilde{Y}_{mis}$ the additive residual, which is independent of θ because of normality. Consequently, the link $\tilde{Y}_{mis} \to \theta$ is completely broken, yielding i.i.d. draws. On the other hand, because $\mathrm{E}[\theta|Y_{mis}, Y_{obs}] = rY_{mis}$, taking $c = r^{-1}$ in $\tilde{Y}_{mis} = Y_{mis} - c\theta = -c(\theta - c^{-1}Y_{mis})$ is equivalent to setting $\tilde{Y}_{mis} = \theta - rY_{mis}$, which is the IRA. The joint normality ensures that $\tilde{Y}_{mis}$ is independent of Y_{mis}, and hence IRA completely breaks the indirect link $Y_{mis} \to \tilde{Y}_{mis}$, again resulting in i.i.d. draws.

2.2 A Phase Transition Phenomenon

In real applications, rarely can the direct or indirect link be broken completely. Even under the normality assumption, we may not be able to compute the regression slopes with infinite precision. A natural question then arises: What happens if we use a c that approximates a regression slope (i.e., from regressing Y_{mis} on θ or θ on Y_{mis})? Does it still retain approximately the benefit of residual augmentation? Common wisdom would suggest so, based on the usual continuity argument.

Unfortunately, the continuity argument would fail here. A clue is offered by considering what happens when $c = 1$, which corresponds to using ASIS for this model (see Yu and Meng, 2011), and when r approaches 1. On the one hand, when $c = 1$, it is easy to see from (2.3) that

$$r_{1\&2} = \frac{r(1-r)}{2} \le \frac{1}{8} \tag{2.4}$$

for all $0 \le r < 1$. On the other hand, for any $c \ne 1$, if we let $r \to 1$, $r_{1\&2}$ will approach 1. Clearly therefore there is a discontinuity at $c = r = 1$. More interestingly or even magically, as proved in the Appendix, the 1/8 bound in (2.4) holds whenever c falls between the two regression slopes, that is, whenever $r \le c \le r^{-1}$, with the bound 1/8 achieved if and only if $r = 1/2$ and $c = 1$.

However, as seen in the perspective plot Figure 3 and the contour plot Figure 4, as soon as c leaves this "safe" zone $[r, r^{-1}]$, the convergence rate $r_{1\&2}$—as a function of (c, r) denoted by $g(c, r)$—increases dramatically, exhibiting essentially a phase transition type of phenomenon at the two boundaries $c = r$ and $c = r^{-1}$. As hinted previously, this phenomenon is most extreme at the point $(c, r) = (1, 1)$: If we fix $c = 1$, then $g(c, r) = r(1-r)/2 \to 0$ as $r \to 1$; if we fix $r = 1$, then $g(c, r) = 1$ for any c (including $c = 1$ by a limiting argument).

A geometric interpretation of this phenomenon can help us to understand it better. The joint (degenerate) normality of $(\theta, Y_{mis}, \tilde{Y}_{mis})$ allows us to visualize the three pairwise (maximal) correlations in a *single* triangle, as in Figure 5, where each vector represents a random variable, and the cosine of the

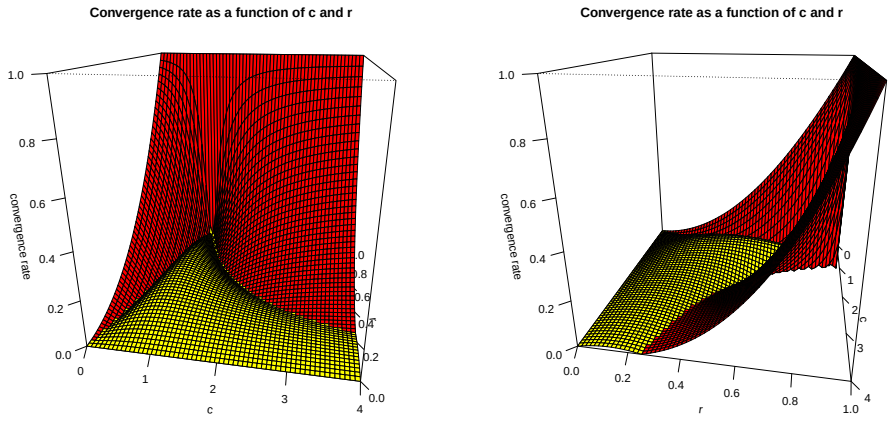

Figure 3: Convergence rate as a function of (c, r) viewed in two perspectives: the yellow (light) area is the "safe" zone, where the convergence rate is bounded by $\frac{1}{8}$; the red (dark) area is outside the two regression lines, where the convergence rate increases dramatically.

(directional) angle between two vectors is their correlation. Denote the pairwise correlations between (θ, Y_{mis}), $(\theta, \tilde{Y}_{mis})$ and $(Y_{mis}, \tilde{Y}_{mis})$ as $\cos\alpha_1 (> 0)$, $\cos\alpha_2$ and $\cos\alpha_3$ respectively. From geometry, we know that $\alpha_2 = \alpha_1 + \alpha_3$. The convergence rate of the interweaving strategy is

$$r_{1\&2} = |\cos\alpha_1 \cos\alpha_2 \cos\alpha_3| = |\cos\alpha_1 \cos(\pi - (\alpha_1 + \alpha_3)) \cos\alpha_3|. \tag{2.5}$$

For a nonobtuse triangle, the product of cosines of its three angles cannot exceed 8^{-1}, hence the same bound is achieved when $\tilde{Y}_{mis}$ falls in the shaded area. Moreover, within the "safe" zone,

$$\mathrm{Corr}(\tilde{Y}_{mis}, Y_{mis})\mathrm{Corr}(\tilde{Y}_{mis}, \theta) \leq 0.$$

This says that the pairwise correlations of $(\theta, \tilde{Y}_{mis})$ and $(Y_{mis}, \tilde{Y}_{mis})$ should have opposite signs to make the interweaving algorithm stable. This finding is consistent with empirical observations and heuristic arguments reported in Yu and Meng (2011) that the interweaving strategy works by taking advantage of the "beauty and beast" nature of two competing DAs. It may also help us search for similar "safe" interweaving algorithms for more complicated problems.

2.3 Going Beyond Normality

But before one conjectures generalizations inspired by this simple example, one must contemplate the possibility that, without the normality condition, such a "safe zone" may completely disappear. After all, the aforementioned 1/8 bound for $r_{1\&2}$ depends critically on the triangulation formulation in (2.5), which was possible because maximal correlation is the same as linear correlation (when it is non-negative) under joint normality. One therefore should at least show such a "safe zone" exists beyond the normality setting.

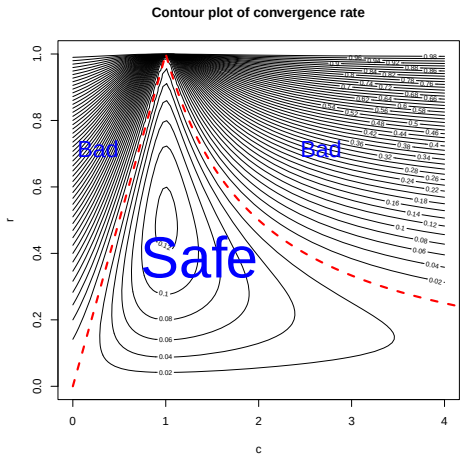

Figure 4: Contour plot: the dashed lines correspond to the two regression slopes, $c = r$ and $c = r^{-1}$.

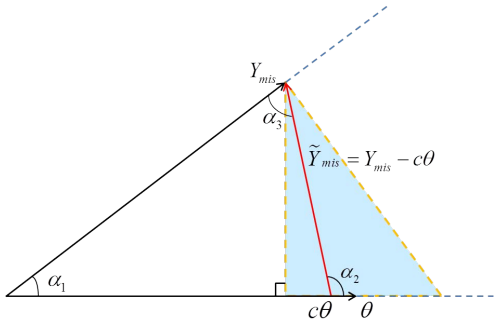

Figure 5: Geometric interpretation: the shaded area corresponds to the "safe" zone, where the formed triangle is nonobtuse.

A common generalization moving beyond normality is to consider a t-type of distribution. Here we consider a general class of the so-called "normal/independent" distributions, which includes the t distribution as a special case (see Lange and Sinsheimer, 1993). This class of (univariate or multivariate) distributions model a random variable Y as $Y = Z/W$ (modulo an affine transformation), where Z is (multivariate) normal, and W is univariate and is independent of Z (and hence the "normal/independent" nomenclature). Obviously, choosing $W = \sqrt{\chi^2_v/v}$ gives the t distribution with v degrees of freedom.

With this setup, let us replace the normal model (2.1) by the following conditional normal model. That is, conditioning on a common variable W, the posterior distribution of (θ, Y_{mis}) is:

$$\begin{pmatrix} \theta \\ Y_{mis} \end{pmatrix} \Bigg| Y_{obs}, W \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{1}{W^2} \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} \right], \tag{2.6}$$

where $0 \le r < 1$ is known and is free of W but may depend on Y_{obs}. The working parameter remains the same, that is, $\tilde{Y}_{mis} = Y_{mis} - c\theta$, and hence (2.2) remains as well other than adding the conditioning on W and the corresponding multiplicative factor W^{-2} for its covariance matrix. Furthermore, the regression slopes remain the same because

$$E[Y_{mis}|\theta, Y_{obs}] = \mathrm{E}\left\{\mathrm{E}[Y_{mis}|\theta, W, Y_{obs}]\,|\theta, Y_{obs}\right\} = \mathrm{E}\left\{r\theta\,|\theta, Y_{obs}\right\} = r\theta$$

and similarly for $\mathrm{E}[\theta|Y_{mis}, Y_{obs}] = rY_{mis}$.

Without restricting the (posterior) distribution of W, we consider in general the maximal correlation between θ and Y_{mis}, which governs the rate of convergence for the DA algorithm for (θ, Y_{mis}). Intuitively, this maximal correlation is determined by two separate sources of dependence, namely the maximal correlation brought in by the common variable W and correlation r between the normal components

after conditioning on W. Mathematically, this intuition is roughly captured by Lemma 1 of Yu and Meng (2011), which in the current case allows us to establish that

$$\mathcal{R}(\theta, Y_{mis}) \leq r + (1-r)\mathcal{R}(\theta, W)\mathcal{R}(Y_{mis}, W). \tag{2.7}$$

Using this inequality together with (1.3) and the fact that $\mathcal{R}(\theta, W) = \mathcal{R}(Y_{mis}, W)$, we can show (see Appendix) that under (2.6), the rate of convergence for interweaving Y_{mis} and $\tilde{Y}_{mis} = Y_{mis} - c\theta$ satisfies

$$r_{1\&2} \leq [g + (1-g)r][g + (1-g)r_1][g + (1-g)r_2], \tag{2.8}$$

where $g = \mathcal{R}^2(\theta, W)$, and

$$r_1 = \frac{|c - r|}{\sqrt{1 + c^2 - 2cr}} \quad \text{and} \quad r_2 = \frac{|1 - cr|}{\sqrt{1 + c^2 - 2cr}}. \tag{2.9}$$

This leads to the "safe" zone $c \in [r, r^{-1}]$ because within this zone, as shown in the Appendix,

$$r_{1\&2} \leq \frac{1}{8}[1 + g]^3, \tag{2.10}$$

which is strictly less than 1 as long as $g = \mathcal{R}^2(\theta, W) < 1$. Note the bound in (2.10) again is independent of the value of r, and is predetermined by the maximal correlation between a normal/independent variable Z/W and its denominator W.

To see how useful the bound in (2.10) can be, let us consider the (bivariate) t distribution, where $W^2 \sim \chi^2_v/v$. Then $\mathcal{R}(\theta, W)$ is simply the maximal correlation between a t random variable and its denominator, which depends only on the degrees of freedom v. We therefore denote it as $\mathcal{R}_v(\theta, W)$ to emphasize this dependence. The analytic calculation of $\mathcal{R}_v(\theta, W)$ seems intractable, but nevertheless we can show that (see Appendix): as $v \to 0$, $\mathcal{R}_v(\theta, W) \to 1$; and as $v \to \infty$, $\mathcal{R}_v(\theta, W) \to 0$. (Incidentally and somewhat ironically, the proof of the latter assertion turns out to be surprisingly difficult, but we were able to establish it by employing a set of well-known theoretical tools for bounding MCMC convergence rate itself.) Therefore (2.10) is a generalization of the 8^{-1} bound under normality because, as $v \to \infty$, the t distribution converges to normal, and $\frac{1}{8}[1 + \mathcal{R}^2_v(\theta, W)]^3 \to \frac{1}{8}$.

For an arbitrary degrees of freedom v, we generated 100,000 t samples and then used the ACE algorithm of Breiman and Friedman (1985)—as given in the *R*-package *acepack*—to estimate the maximal correlation $\mathcal{R}_v(\theta, W)$. For a given v, this process was repeated 50 times to construct (95%) confidence intervals, represented by the "vertical dots" in the left panel of Figure 6, which plots the resulting estimated curve of $\mathcal{R}_v(\theta, W)$ as a function of v (on a equal-spaced grid 0 to 10 plus $v = 20$). The right panel plots the corresponding bound in (2.10), using the $g = \mathcal{R}^2_v(\theta, W)$ values displayed in the left panel.

We see from the right panel that as soon as $v \geq 6$, the rate appears to not exceed 1/5. Even for $v = 1$, that is, the Cauchy distribution, the *upper confidence limit* on the upper bound of the convergence rate is only about 1/2. Whereas these bounds are not as good as 1/8 for the normal case, they are far better

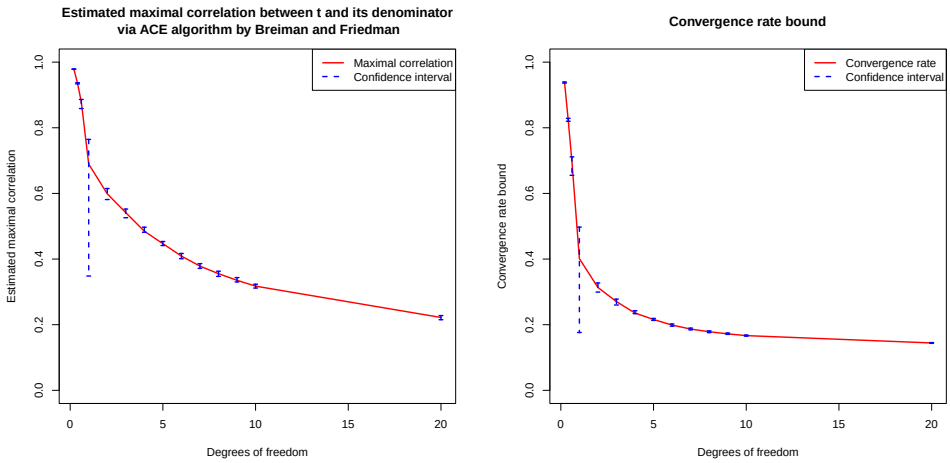

Figure 6: Estimated $\mathcal{R}_v(\theta, W)$ and convergence rate bound: the left plot is the estimated $\mathcal{R}_v(\theta, W)$ via ACE; the right plot is the corresponding value given by (2.10), an upper bound for convergence rate.

than adequate for practical purposes, considering most numerical bounds used in practice are above 0.9 and even above 0.99; see for example Hobert (2001a) and van Dyk and Meng (2001, Rejoinder). Indeed, if we use 0.9 as standard, then unless one fits a t model with tiny fractional degrees of freedom (e.g., $v \leq 0.1$), the interweaving algorithm will be safe as long as $c \in [r, r^{-1}]$.

Regarding the phase transition phenomenon, for the normal model (2.1) we were able to demonstrate it exactly because the chain was reversible and the inequality (1.3) becomes equality under that normal model. For this more general normal/independence model, we currently can only demonstrate such a phenomenon for the bound in (2.8). This is given in Figure 7, where the four values of g correspond to four values of the degrees of freedom v in the left panel of Figure 6. We see clearly the very similar shape as in Figure 3, other than that the function values in the safe zone increase as g increases. This of course only provides suggestive evidence (and it is only for t distributions), and we certainly hope a more direct demonstration can be found, perhaps via finding a lower bound that shares a similar shape as in Figure 7.

Regardless of the extent to which the phase transition phenomenon exists for an arbitrary choice of W, it is clear that the choice of $c = 1$ is always safe irrespective of the actual value of r (which depends on the actual data) in the sense that (2.10) always hold for $c = 1$. Indeed, here $c = 1$ can be viewed as a minimax choice because it minimizes the maximal convergence rate (strictly speaking, an upper bound of the rate) against all possible values of r.

A couple of remarks are in order before we illustrate the power of applying the interweaving strategy with residual augmentations in a practical setting. First, the demonstrations above are to provide theoretical insights (e.g., the phase transition phenomenon) and to illustrate theoretical potential (e.g., the upper bound (2.10)); they do not take into account the issue of ease or cost of implementation, an

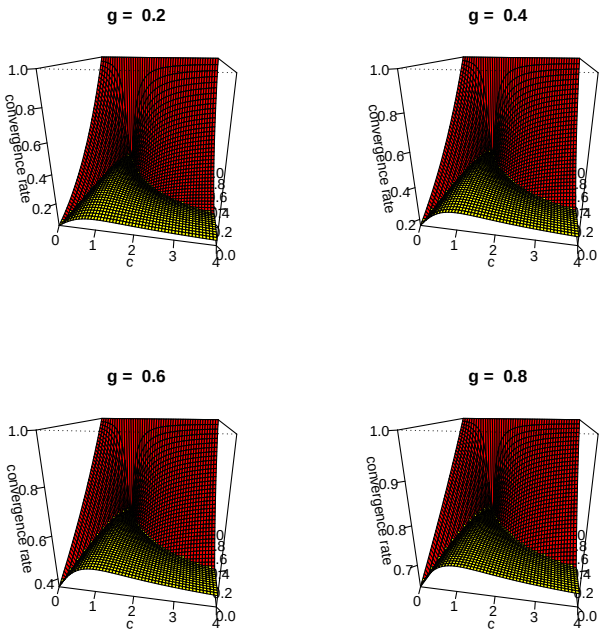

Figure 7: Bound in (2.8) as a function of (c, r), with four values of $g = \mathcal{R}^2(Z/W, W)$. Note the different ranges of values on the vertical axis.

issue that will be investigated in the next section. Second, as a side note, the very wide confidence intervals seen in Figure 6 only at $v = 1$ may seem puzzling at the first sight, because one might expect the Monte Carlo error getting progressively worse as v decreases below 1, which corresponds to tail behavior that is even heavier than Cauchy (so much so that a sample mean is more dispersed than a single observation). Whereas we do not have a good explanation for this phenomenon, we suspect it is related to a hidden symmetry in the maximal correlation, that is, $\mathcal{R}(Z/W, W) = \mathcal{R}(W/Z, W)$, with the Cauchy distribution corresponding to the center of symmetry because it is invariant to the reciprocal transformation, and hence its unique properties.

3 An Empirical Exploration via Probit Regression

3.1 A Locally Linearized Direct Residual Augmentation

Consider the widely used Probit regression model:

$$Y_{obs,i} = \text{sign}(Y_{mis,i}), \qquad Y_{mis,i}|\theta, X_i \sim N(X_i\theta, 1), \tag{3.1}$$

where $Y_{obs,i}$ is the observed binary (± 1) outcome, the sign of the latent score $Y_{mis,i}$, X_i is a $1 \times p$ vector of covariates, and θ is a $p \times 1$ vector of regression coefficients. Write $Y_{obs} = (Y_{obs,1}, \cdots, Y_{obs,n})^\top$,

$Y_{mis} = (Y_{mis,1}, \cdots, Y_{mis,n})^\top$, and $X^\top = (X_1^\top, \cdots, X_n^\top)$. Taking the standard noninformative prior $p(\theta) \propto 1$, we have the well-known full conditional distributions for the standard DA/Gibbs sampler (see Albert, 1992; Albert and Chib, 1992; Meng and Schilling, 1996):

$$Y_{mis,i}|Y_{obs}, \theta \sim \text{TN}(X_i\theta, 1, Y_{obs,i}); \tag{3.2}$$

$$\theta|Y_{obs}, Y_{mis} \sim N(\hat{\theta}, (X^\top X)^{-1}). \tag{3.3}$$

Here $\hat{\theta} = (X^\top X)^{-1} X^\top Y_{mis}$, and $\text{TN}(\mu, \sigma^2, Y_{obs,i})$ denotes a $N(\mu, \sigma^2)$ distribution truncated to the interval $(0, \infty)$ if $Y_{obs,i} = 1$ and to $(-\infty, 0)$ if $Y_{obs,i} = -1$. The standard DA/Gibbs sampler iterates between (3.2) and (3.3). Though convenient, it can be extremely slow. Several methods, therefore, have been proposed to improve it, including PX-DA (Liu and Wu, 1999) and ASIS (Yu and Meng, 2011). Below we first demonstrate how to implement the residual augmentation, and then we compare it to several existing algorithms.

Given the normal model in (3.1) for Y_{mis}, which is univariate (in contrast to θ, which is often multivariate), it is easier to consider the additive DRA $\tilde{Y}_{mis} = Y_{mis} - \text{E}[Y_{mis}|\theta, Y_{obs}]$. In particular, it is known that if we let $H(z) = z + M(z)$ where $M(z) = \frac{\phi(z)}{\Phi(z)}$ is the inverse Mills ratio, then (recall $Y_{obs,i} = \pm 1$)

$$\text{E}[Y_{mis,i}|\theta, Y_{obs}] = \text{E}[Y_{mis,i}|\theta, Y_{obs,i}] = Y_{obs,i} H(Y_{obs,i} X_i \theta), \qquad i = 1, \ldots, n. \tag{3.4}$$

However, since $H(\pm X_i\theta)$ is non-linear in θ, deriving and then sampling from the resulting $p(\theta|Y_{obs}, \tilde{Y}_{mis})$ is rather a difficult task. As a compromise, we seek a locally linear approximation to $H(z)$ by utilizing its derivative

$$G(z) = H'(z) = 1 - zM(z) - M^2(z).$$

The resulting local residual augmentation has the form

$$\tilde{Y}_{mis,i} = Y_{mis,i} - G(Y_{obs,i} X_i \theta) X_i \theta, \qquad i = 1, \ldots, n. \tag{3.5}$$

However, the corresponding $p(\theta|Y_{obs}, \tilde{Y}_{mis})$ is still hard to employ, because of the θ inside the non-linear $G(\cdot)$ function.

3.2 Seeking Compromise via Adaptive Data-Dependent Augmentation

To further simplify the implementation, we adopt the adaptive MCMC idea (see Rosenthal, 2011 and references therein). That is, at $(t+1)$st iteration, we adopt a DA scheme that depends on the value of $\theta^{(t)}$:

$$\tilde{Y}_{mis,i} = Y_{mis,i} - b_i^{(t)} X_i \theta, \quad \text{where} \quad b_i^{(t)} = G(Y_{obs,i} X_i \theta^{(t)}). \tag{3.6}$$

It is critical to recognize that (3.5) and (3.6) are different augmentation schemes, even though they share the same conditional distribution $p(\tilde{Y}_{mis}^{(t+1)}|\theta = \theta^{(t)}, Y_{obs})$. Their difference lies in the two different

conditional distributions $p(\theta^{(t+1)}|\tilde{Y}_{mis} = \tilde{Y}_{mis}^{(t+1)}, Y_{obs})$. For scheme (3.5), the θ inside the $G(\cdot)$ function is free or "live", therefore we need to take it into account when deriving $p(\theta|\tilde{Y}_{mis}; Y_{obs})$ for drawing $\theta^{(t+1)}$. In contrast, for scheme (3.6), the $\theta^{(t)}$ inside the $G(\cdot)$ is fixed or "dead" by iteration $(t+1)$st, so in deriving $p(\theta|\tilde{Y}_{mis}; Y_{obs})$, $b_i^{(t)} = G(Y_{obs,i}X_i\theta^{(t)})$ is just a constant, rending $\tilde{Y}_{mis}$ of (3.6) truly linear in θ.

This "adaptive linearity" on one hand permits an easy implementation, but on the other hand destroys the proper convergence of the resulting Markov chain. This is because the adaptive DA, namely an iteration-dependent DA model $p^{(t)}(\theta, Y_{mis}|Y_{obs})$, can easily destroy the detailed balance condition. Whereas the detailed balance condition is not necessary for MCMC to converge, without it proper convergence can be easily destroyed. As a matter of fact, our empirical checking indicated that our adaptive algorithm does not converge to our desired target, as demonstrated in Figure 10 of Section 3.5 below.

Fortunately this is a relatively easy problem to resolve, because the reason we invoke the adaptation is to seek a suitable compromise between simplicity and speed. We therefore can run the adaptive algorithm for a burn-in period, say until $t = t^*$, and then fix $b_i^{(t)} = b_i$ for all $t > t^*$ (and all i), eliminating adaptation. Here the value b_i can be chosen in many ways by analyzing $\{b_i^{(t)}, t \le t^*\}$, such as the average of the last (say) 10% of the $\{b_i^{(t)}, t \le t^*\}$. Another way to motivate this switching strategy is to consider the adaption as a greedy strategy, i.e., it aims to find the best piece-wise linear approximation given the θ drawn at the current iteration. But what we really need is a good approximation given θ within a reasonable range as determined by its posterior distribution. Therefore, at the end of the adaptive stage, we form a compromise by taking an appropriate summary of b_i's from the adaptive stage. Currently we do not have a general theoretical framework for choosing the optimal summary. Nor do we believe there is a unique optimal choice here, because such a choice typically entails a trade-off between statistical efficiency and computational efficiency. Nevertheless, we conducted a preliminary empirical investigation of the impact of the choices of b_i, as reported is in Section 3.5 below.

In contrast to a global residual augmentation such as $\tilde{Y}_{mis} = Y_{mis} - cX\theta$, where c is a scalar working parameter, the adaptation outlined above allows us to search for a more powerful residual augmentation (for our goal to reduce auto-correlations) by taking into account heterogeneity in different components as governed by the actual observed data. Specifically, the adaption leads to a component-wise (direct) residual augmentation in the form of

$$\tilde{Y}_{mis} \equiv \begin{pmatrix} \tilde{Y}_{mis,1} \\ \tilde{Y}_{mis,2} \\ \vdots \\ \tilde{Y}_{mis,n} \end{pmatrix} = \begin{pmatrix} Y_{mis,1} - b_1X_1\theta \\ Y_{mis,2} - b_2X_2\theta \\ \vdots \\ Y_{mis,n} - b_nX_n\theta \end{pmatrix} = Y_{mis} - BX\theta, \tag{3.7}$$

where $B = \text{diag}\{b_1, \ldots, b_n\}$. What makes (3.7) more powerful than $\tilde{Y}_{mis} = Y_{mis} - cX\theta$ is not only that it permits heterogeneity among the n components, but more importantly the value of individual working parameter b_i *takes into account the information from the actual observed data* because it depends on the

value of $Y_{obs,i}$ as seen in (3.6). This is an important extension of virtually all previous data augmentation schemes, which—to the best of our knowledge— were constructed before seeing the actual data, at least for routine applications such a probit regression. But see Section 4 for a discussion of the potential robust-efficiency trade-off from using data-dependent augmentation schemes.

3.3 A Prototype Algorithm

With the setup outlined above, we can carry out (at least) two algorithms. The first is simply a direct DA algorithm using (3.7) as its augmentation scheme, albeit we need to deal with its adaptive nature, as outlined below. The second is to interweave the first with the standard DA based on the original DA Y_{mis} to gain additional benefit. Below we provide the details for the first, as the interweaving one is rather trivial once the first one is in place.

Specifically, the direct (initially) adaptive DA algorithm requires two-stage execution:

I. *Adaptive Stage*: $t = 1, \ldots, t^*$, update $b_i = b_i^{(t)} (i = 1, \ldots, n)$ at each iteration;

II. *Sampling Stage*: Same as Adaptive Stage, except b_i is fixed as $\bar{b}_i$, the average of the last 10% of the $b_i^{(t)}$'s obtained from the Adaptive Stage $(i = 1, \ldots, n)$. (See Section 3.5 for other choices.)

Operationally, during the Adaptive Stage, we carry out the following (where $\tilde{Y}_{mis} = (\tilde{Y}_{mis,1}, \ldots, \tilde{Y}_{mis,n})^\top$):

- Draw $\tilde{Y}_{mis}^{(t+1)} | \theta^{(t)}, Y_{obs}$:

Step 1 Update $b_i \Leftarrow b_i^{(t)} = G(Y_{obs,i} X_i \theta^{(t)}), \quad i = 1, \ldots, n;$

Step 2 Draw $Y_{mis,i}^{(t+1)} \sim \mathrm{TN}(X_i \theta^{(t)}, 1, Y_{obs,i})$ and then compute $\tilde{Y}_{mis,i}^{(t+1)} = Y_{mis,i}^{(t+1)} - b_i X_i \theta^{(t)}$.

- Draw $\theta^{(t+1)} | \tilde{Y}_{mis}^{(t+1)}, Y_{obs}$:

Step 3 For $i = 1, \ldots, n$, compute $\tilde{X}_i = (1 - b_i) X_i$ and then

$$\hat{\mu} = (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top \tilde{Y}_{mis}^{(t+1)}, \quad \hat{\Sigma} = (\tilde{X}^\top \tilde{X})^{-1},$$

where $\tilde{X}^\top = (\tilde{X}_1^\top, \ldots, \tilde{X}_n^\top)$;

Step 4 Draw $\theta^{(t+1)} | \tilde{Y}_{mis}^{(t+1)}, Y_{obs} \sim \mathrm{TN}(\hat{\mu}, \hat{\Sigma})$ with the truncation determined by the constraint that $\mathrm{sign}(\tilde{Y}_{mis,i}^{(t+1)} + b_i X_i \theta^{(t+1)}) = Y_{obs,i}$. We implement this step via a nested Gibbs sampler: for each i, draw $\theta_i^{(t+1)} | \tilde{Y}_{mis}^{(t+1)}, Y_{obs}, \theta_{-i}^{(t+1)}$, a truncated univariate normal distribution, and repeat it K cycles.

At the Sampling Stage, we simply skip Step 1, that is, we use $b_i = \bar{b}_i$ for all iterations to produce our MCMC samples. We emphasize that this algorithm is by no means optimal in any sense; there should be many ways to improve upon it especially regarding the potentially time consuming nested Gibbs sampler used in Step 4; see Section 3.5 for an exploration. Our aim here is to provide the first

prototype algorithm, in a real application setting, that builds upon the concept of residual augmentation formulated in Yu and Meng (2011). Nevertheless, our preliminary numerical experiments, as reported below, have shown great potential even for this prototype algorithm.

3.4 A Numerical Comparison

To see the effectiveness of our prototype algorithm and of its interweaving with the standard Gibbs sampler, we conducted a numerical experiment using the lupus nephritis data set of van Dyk and Meng (2001; Table 1), which has $n = 55$ patients and $p = 3$ covariates (including a constant term for the intercept). We compare it with various other algorithms. The algorithms we included in our comparisons are:

I. Standard Gibbs sampler given by (3.2)-(3.3). This is also known as the DA algorithm with *Sufficient Augmentation* (SA) Y_{mis} (Yu and Meng, 2011), and hence it is the same as setting $b_i \equiv 0$ in our prototype algorithm for all i (therefore $\tilde{Y}_{mis} = Y_{mis}$, which makes Step 4 the same as (3.3)).

II. A marginal augmentation/PX-DA algorithm based on a *multiplicative* working parameter $\tilde{Y}_{mis} = \sigma Y_{mis}$, with Haar working prior $p(\sigma^2) \propto \sigma^{-2}$—see Liu and Wu (1999) and van Dyk and Meng (1999).

III. The DA algorithm based on *Ancillary Augmentation* (AA) $\tilde{Y}_{mis} = Y_{mis} - X\theta$ (Yu and Meng, 2011), which is the same as setting $b_i \equiv 1$ in our prototype algorithm for all i.

IV. The ASIS algorithm (Yu and Meng, 2011) that interweaves SA and AA in (I) and (III) respectively.

V. Our prototype DRA algorithm as given in Section 3.3.

VI. The algorithm that interweaves (I) and (V) (IS-DRA).

Here the first two are well-known algorithms in the literature, which we employ as benchmarks, even though algorithm II uses a multiplicative working parameter (and hence theoretically it is not directly comparable with those built upon an additive working parameter). The next two are more recent algorithms proposed in Yu and Meng (2011) but without the benefit of tuning b_i according to the actual data because they set $b_i = 1$ for all i. The last two are our new prototype algorithms (without and with interweaving), benefiting from allowing data to have a strong influence on choosing b_i $(i = 1, \ldots, n)$. Figure 8 displays the trajectories, histograms, and autocorrelations of the draws of θ_1 (the first coefficient) for these six algorithms. We see the algorithms work progressively better, at least in terms of autocorrelations, empirically demonstrating that we are on the right track in our effort to reduce autocorrelations by using progressively more efficient DA schemes and with the help of the interweaving strategy.

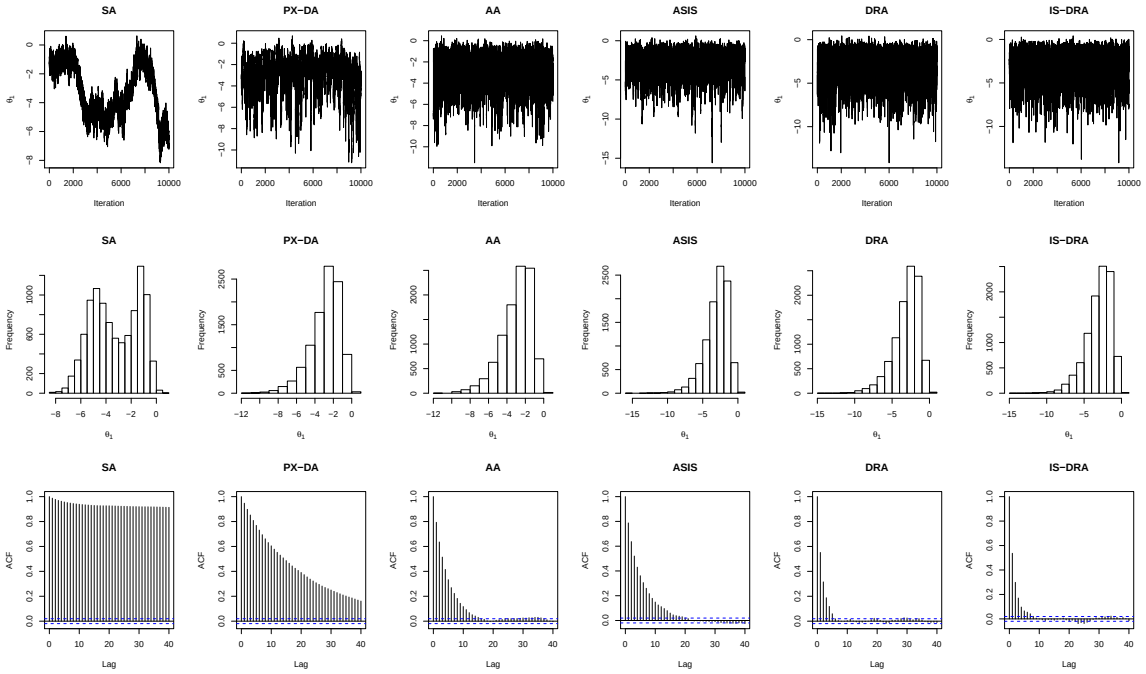

Figure 8: Comparing various samplers for the lupus nephritis data: trajectories, histograms and autocorrelations ($K = 30$ for III-VI).

Clearly autocorrelation measures only (one aspect of the) statistical efficiency. Another important measure is CPU time, an aspect of computational efficiency. Table 1 reports the CPU time (in seconds) from 25 replications, together with estimated effective sample size (ESS) obtained from the *R*-package *coda*. We see that in terms of both ESS and Relative Speed, DRA(30)—with or without IS— ranks at the top, about one order of magnitude improvement over PX-DA and about two orders of improvement over SA. However, unlike the statistical efficiency measure ESS, which does not depend on how the algorithm is actually implemented (as long as it is implemented correctly), CPU time depends critically on how the algorithm is implemented, in what program language(s) it is written, on what machine it is carried out, etc. As a matter of fact, when we initially implemented Step 4 directly via the *R* language, the drawing from the truncated multivariate normal turned out to be so costly that the gain in statistical efficiency by DRA was outweighed by the lost of computational efficiency. The results in Table 1 are from using the *R*-package *tmvtnorm* (by Stefan Wilhelm), which was implemented in Fortran. All the rest of the implementation was done in *R*, except for the drawing from a truncated multivariate uniform as needed by AA and hence also by ASIS (which corresponds to Step 4 of our prototype algorithm; see Section 4.1 of Yu and Meng, 2011). These truncated uniform drawings were also carried out by the same Fortran program *tmvtnorm* with the covariance matrix set to a very large value, so the truncated multivariate normal effectively becomes truncated multivariate uniform. We adopted this strategy to

Method	Mean Time	ESS (min,median,max)	ESS/Time	Relative Speed
SA	20.2	(4, 16, 81)	0.8	1
PX-DA	20.6	(180, 235, 273)	11.4	14
AA(1)	19.2	(64, 115, 157)	6	7
AA(10)	22.5	(211, 454, 601)	20.2	25
AA(30)	27.9	(871, 1025, 1235)	36.8	46
ASIS(1)	24.9	(78, 122, 203)	4.9	6
ASIS(10)	27.3	(363, 475, 592)	17.4	22
ASIS(30)	32.8	(771, 1047, 1337)	32	40
DRA(1)	20.2	(199, 259, 366)	12.8	16
DRA(10)	22.9	(976, 1233, 1458)	53.8	67
DRA(30)	28.8	(2216, 2928, 3968)	101.6	127
IS-DRA(1)	24.7	(222, 285, 356)	11.6	14
IS-DRA(10)	27.5	(1033, 1294, 1503)	47.1	59
IS-DRA(30)	33.3	(2363, 2950, 3286)	88.6	110

Table 1: Comparison of methods after $10,000$ samples averaged over 25 runs. For AA, ASIS, DRA and IS-DRA, the number within the parentheses (e.g. DRA(K)) denotes the number of iterations in the nested Gibbs sampler. For *relative speed*, we use SA as the reference point, e.g., in this application, IS-DRA(30) is about 110 times faster than SA in terms of the relative speed.

ensure a meaningful comparison of CPU times so the simulation results do not bias toward our DRA; indeed, when we implemented AA and ASIS completely in R, their CPU time was much worse than our DRA using *tmvtnorm*, further illustrating how computational efficiency depends critically on the actual implementation, not just the algorithm itself.

We remark here that the substantial increases in ESS as K increases from $K = 1$ to $K = 30$ clearly demonstrate the importance and effectiveness of using data augmentation schemes that are as close to residual augmentations as possible. We also note that in this case the additional gain/protection from using interweaving is rather minor, a consequence of a rather effective DRA for this problem and particular data set. Yu and Meng (2011) provided ample evidence that the performance of any single DA tends to depend on the actual data set much more substantially than those by interweaving a pair. Our theoretical bounds given in Section 3 (albeit they do not apply to the Probit regression problem) provide further suggestive evidence of the robust nature of our interweaving strategy.

3.5 Seeking Effective Data-Dependent Working Parameter

In Section 3.2 we emphasized the importance and potential of allowing the actual data to govern the choice of the working parameter. In the current cases, the working parameters are $\{b_i, i = 1, \ldots, n\}$.

In Section 3.3 we then mentioned that there are a number of possible choices of b_i for the sampling stage based on working parameter values obtained during the adaptive stage: $\{b_i^{(t)}, t = 1, \ldots, t^*\}$. As a preliminary assessment of the impact of the choice of data-dependent working parameters on ESS, Figure 9 displays the box-plots of ESS for six choices of b_i's. They are:

1. Last: Set $b_i = b_i^{(t*)}$, the last updated value of b_i from the adaptive stage;
2. Mean: Set $b_i = \bar{b}_i$, the average of the last 10% $b_i^{(t)}$'s from the adaptive stage;
3. Median: Set $b_i = \text{med}\{b_i\}$, the median of the last 10% $b_i^{(t)}$'s from the adaptive stage;
4. Mode: Set $b_i = \text{mode}\{b_i\}$, the mode of an estimated density (using a kernel method) of the last 10% $b_i^{(t)}$'s from the adaptive stage;
5. Mode2: Set $b_i = G(Y_{obs,i} X_i \hat{\theta})$ (see (3.6)), where $\hat{\theta}$ is the mode of an estimated density (using a kernel method) of the last 10% of $\theta^{(t)}$'s from the adaptive stage;
6. MLE: Set $b_i = G(Y_{obs,i} X_i \hat{\theta}_{MLE})$, where $\hat{\theta}_{MLE}$ is the MLE of θ under the Probit model. (This last choice of b_i does not require the adaptive stage, and it is included as a benchmark.)

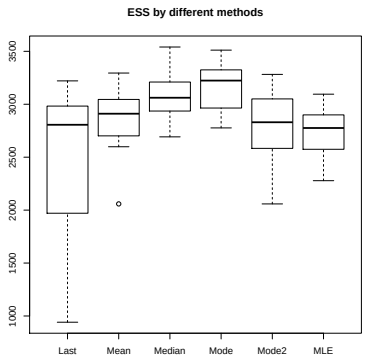

Figure 9: Comparing various choices of the data-dependent working parameter based on 25 simulations.

As one may expect, using only the last value from the adaptive stage creates too much variation, although it is the least costly in terms of CPU time and easiest to implement. The latter is true for using the MLE, which eliminates the adaptive stage altogether. Unfortunately these two methods also have the lowest ESS on average. The best performer seems to be using the mode of $b_i^{(t)}$, but it involves a kernel estimation (and mode finding), which can be more costly timewise, albeit for our simulation this was not a problem. For our Table 1, we adopted the mean choice because we conjectured that it would represent a practical compromise between statistical efficiency and computational efficiency. Figure 9 suggests, however, that the median perhaps is an even better compromise. Further research obviously is needed to assess whether using the median (or another method) is a good compromise in general.

We did, however, find a partial indication for the better performance of the median when we attempted to address both reviewers' question about how we could trust b_i's from the adaptive stage, where the draws from our prototype algorithm itself cannot be trusted. The answer lies in the fact that we are not seeking the theoretically optimal choice of b_i, but rather any reasonable choice of it that would result in an algorithm with acceptably satisfactory efficiency. Recall the choice of b_i does not affect the validity of our prototype algorithm as long as it is fixed during the sampling stage. Furthermore, although in the adaptive stage the draws of θ from our algorithm follow a different distribution than the one for the draws from the sampling stage, the two distributions apparently are close enough that their corresponding distributions for $b_i = G(Y_{obs,i}X_i\theta)$ do not provide significantly different summary statistics, especially for the robust ones such as medians.

To illustrate this point, Figure 10 displays the Q-Q plots of the samples of the three components of θ from the sampling stage against their counterparts from the adaptive stages. We see clearly that although the plots show some visible differences between the two distributions, the differences lie primarily in their tails.

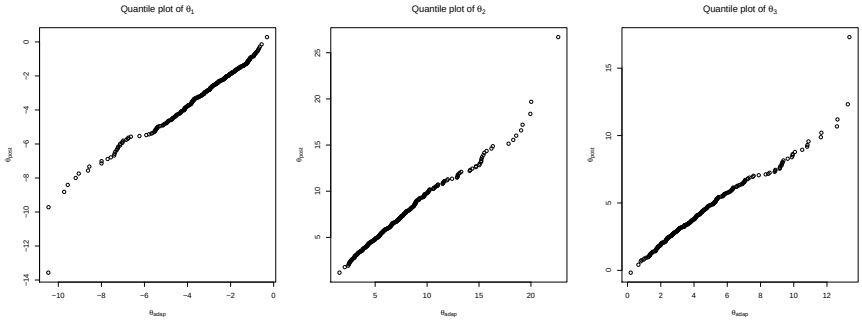

Figure 10: Q-Q plots for samples of $\theta_i, i = 1, 2, 3$: sampling stage verses the adaptive stage.

Figure 11 displays the corresponding Q-Q plot of samples of randomly selected four (out of $n = 55$) b_i from the sampling stage against those from the adaptive stage. Again we see the major differences lie in tails. To see the impact of adaptation numerically, let $\bar{b}_i^{[S]}$ and $\bar{b}_i^{[A]}$ be the sample means of b_i from the sampling stage and the adaptive stage respectively. Then for the same data underlying Figure 10 and Figure 11, we have (recall $n = 55$ for our lupus nephritis data set)

$$\frac{1}{55}\sum_{i=1}^{55}\left|\bar{b}_i^{[S]} - \bar{b}_i^{[A]}\right| = 0.00303, \tag{3.8}$$

which is only one third (0.00303/0.918=0.0033) of a percent compared with $\sum_{i=1}^{55}\bar{b}_i^{[S]}/55 = 0.918$. If we replace the sample means in (3.8) by their sample median counterparts, then the absolute difference will be even smaller: 0.00226. The corresponding relative difference compared with the average of the sample medians is $0.00226/0.926 = 0.0024$, only one quarter of a percent. We therefore have rather good

empirical verification that the lack of proper convergence during the adaptive stage had non-significant impact on our overall findings.

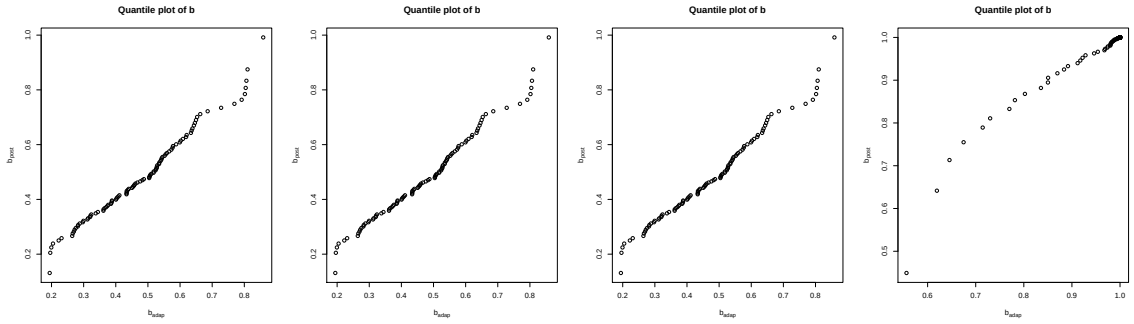

Figure 11: Q-Q plot: samples of b_i (for four different i's) from the sampling and adaptive stages.

It is also worthy to point out the trade-off between the number of iterations during the adaptive stage and the sampling stage. Practically, the length of the adaptive stage is likely to be positively correlated with the quality of our choices of b_i's for the sampling stage. However, the lengths of the two stages obviously compete with each other for a total given computational budget. There is hence a trade-off between a longer sampling stage with a less effective choice of b versus a shorter sampling stage but with a more effective b. We still need to develop a good practical guideline for such a trade-off, as well as for a number of other trade-offs discussed in the next Section.

4 Challenges and Opportunities

The primary purpose of this article was to provide initial evidence of the potential of residual augmentations, proposed recently (Yu and Meng, 2011). On the theoretical side, we demonstrated the possibility of establishing numerically rather tight universal bounds (e.g., 1/8 or 1/5) by utilizing a unique "competing" nature of the interweaving strategy, as well as the existence of the "safe zone" and the resulting convenient minimax choice ASIS. At the same time, we uncovered a somewhat unexpected phase transition behavior, which makes the issue of robustness-efficiency trade-off particularly critical and tricky. Mathematically, the situation reminds us of AR(1) type of time series models (see Pena, Tiao and Tsay, 2001), where the unit root serves as the only boundary between a stationary region and an explosive region. It is well known that uncertainties in identifying a unit root could lead to rather different statistical properties from what were intended (e.g., Meng and Xie, 2013). Analogously errors in computing regression slopes for residual augmentations could mean the difference between delivering nearly independently and identically distributed draws and producing almost identical draws (because of extremely poor mixing)!

Therefore, much needs to be done in order to identify situations where we can push the data-dependent

residual schemes to achieve their maximal efficiency, and where it is too dangerous to do so and hence we should stay with robust "data-free" schemes such as ASIS. Whereas we succeeded in establishing such results for a class of normal/independent models, we nevertheless benefited from the conditional normality inherited in such a class of models and the symmetric nature of (θ, Y_{mis}) as in (2.6). We imagine the task is rather challenging in general because without normality of some sort (marginal or conditional), the analytic manipulation of maximal correlations is typically intractable. Furthermore, the three maximal correlations in (1.3) generally cannot be mapped into the same triangle because (for example) the function (i.e., transformation) of θ that leads to its maximal correlation with Y_{mis} may not be the same function for maximizing its correlation with $\tilde{Y}_{mis}$. This would render the geometric expression (2.5) inapplicable, at least not directly. Nevertheless, given the general difficulties in establishing useful bounds for convergence rates for MCMC (see various chapters in Brooks et al., 2011), we are encouraged by the preliminary theoretical results reported in Section 3.

On the algorithmic side, as we have seen from the Probit models, there are at least two issues we need to deal with effectively in order to fully realize the potential of residual augmentations, with or without interweaving. The first is how to find a good compromise between statistical efficiency, which requires us to stay as closely as possible to the theoretically optimal residuals (under whatever criterion adopted), and implementation/computational efficiency, which often requires simple approximations to the optimal residual for effective execution of the $\tilde{Y}_{mis} \to \theta$ step. The second is that even when we know how to carry out the $\tilde{Y}_{mis} \to \theta$ step in theory, its actual implementation can have a significant impact on the overall competitiveness of the resulting algorithm. As seen in Section 3.4, different implementations of the nested Gibbs sampler have led to very different algorithmic efficiency.

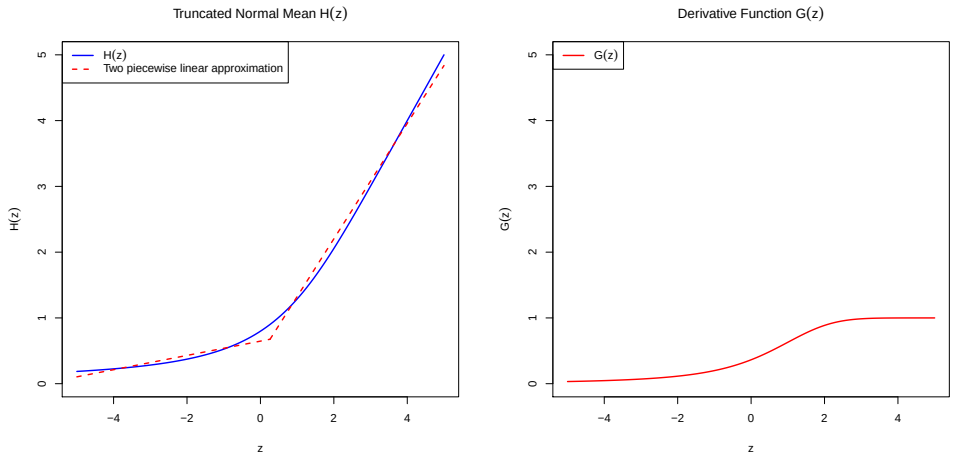

Figure 12: The left plot is the H function with a two-piecewise linear approximation; the right plot is the derivative function $G(z) = H'(z)$.

We are also working on finding a better approximation to the actual residual than our current linear

approximation. As is evident from the left panel in Figure 12, the conditional mean function $H(z)$ can be approximated well by a two-piecewise linear function. That is, we can find two suitable derivative values (see the right panel) $G(z)$ as the b_i's for our residual augmentations $\tilde{Y}_{mis,i} = Y_{mis,i} - b_i X\theta$ depending on the value of $X\theta$ (from the left panel, choosing 0 as the connecting point of the two linear pieces seems to be both effective and convenient). However better approximations do not transfer to better algorithms unless the added computational burden does not unduly offset the gain in statistical/algorithmic efficiency.

Furthermore, for our probit regression we have constructed only DRA, mainly because in this case DRA is simpler in construction than IRA due to the fact that $\mathrm{E}[Y_{mis,i}|\theta, Y_{obs}] = \mathrm{E}[Y_{mis,i}|\theta, Y_{obs,i}]$, permitting component-wise (as indexed by i) calculations (see (3.4)), which is not the case for $\mathrm{E}[\theta|Y_{mis}, Y_{obs}]$. However, for each component i, we notice that the probit model depends on θ only through $X_i\theta$. Hence it is possible to consider forming component-wise IRA in the form of $\tilde{Y}_{mis,i} = X_i\theta - \mathrm{E}[X_i\theta|Y_{mis,i}, Y_{obs,i}]$, which would still render *component-wise zero posterior correlation*:

$$\mathrm{Cov}(\tilde{Y}_{mis,i}, Y_{mis,i}|Y_{obs,i}) = 0, \quad i = 1, \ldots, n. \tag{4.1}$$

Although component-wise derivation/calculation will render implementation simplicity, we are likely giving up some statistical efficiency because (4.1) does not achieve the actual zero posterior correlation $\mathrm{Cov}(\tilde{Y}_{mis}, Y_{mis}|Y_{obs}) = 0$; note (4.1) implies neither $\mathrm{Cov}(\tilde{Y}_{mis,i}, Y_{mis,j}|Y_{obs,i}) = 0$ for any $i \neq j$, nor $\mathrm{Cov}(\tilde{Y}_{mis,i}, Y_{mis,i}|Y_{obs}) = 0$ for any i.

We are currently investigating a number of such trade-off issues between statistical efficiency and computational efficiency (e.g., implementation simplicity). There are many challenges ahead, and what is reported above are only those from our initial investigation. At the same time, we have so many options to explore, from forming DRA and IRA to many of their approximations and variations (e.g., component-wise residuals), and with or without interweaving. In our general pedagogical effort, explaining the difference between regressing Y on X and regressing X on Y, to those who are ingrained in deterministic thinking of a functional relationship, has not been a trivial task. But it is the very existence of these two regression lines that offers us a unified theme to explore and construct MCMC algorithms which come closer to realizing the sweet 3-S dream, a dream we invite all readers to share.

5 Appendix

5.1 Proof of the 8^{-1} bound for the normal model

Proof: Let $g(r, c)$ be the function defined by the right-hand side of (2.3). Then when $r \leq c \leq r^{-1}$,

$$g(r, c) = \frac{r(c-r)(1-rc)}{1+c^2-2rc} \leq \frac{1}{8} \quad \Leftrightarrow \quad (1+8r^2)c^2 - 2r(4r^2+5)c + 1 + 8r^2 \geq 0. \tag{5.1}$$

But for the quadratic form (in c) on the right-hand side, the discriminant $\Delta = 4(2r+1)^2(2r-1)^2(r^2-1) \leq 0$. This establishes our claim that when c is in the safe zone $[r, r^{-1}]$, $r_{1\&2} \leq 8^{-1}$. (As mentioned before,

a geometric proof is to use (2.5). It can also be viewed as a special case of the t model with infinite degrees of freedom, discussed below.)

5.2 Proof of the bound (2.10) for the normal/independence model

To prove this bound we need the notion of *partial maximal correlation* (Yu and Meng 2011) defined for three random variables X, Y, Z as the following:

$$\mathcal{R}_Z(X,Y) = \sup_{f,h\in L^2} \frac{\mathrm{Cov}(f(X)-\mathrm{E}[f(X)|Z], h(Y)-\mathrm{E}[h(Y)|Z])}{\sqrt{\mathrm{Var}(f(X)-\mathrm{E}[f(X)|Z])\mathrm{Var}(h(Y)-\mathrm{E}[h(Y)|Z])}}. \tag{5.2}$$

We also need the notion of *conditional maximal correlation*, which is defined as

$$\begin{aligned}\mathcal{R}(X,Y|Z) &= \sup_{f,h\in L^2} \mathrm{Corr}(f(X),h(Y)|Z)\\ &= \sup_{f,h\in L^2} \frac{\mathrm{Cov}(f(X),h(Y)|Z)}{\sqrt{\mathrm{Var}(f(X)|Z)\mathrm{Var}(h(Y)|Z)}}.\end{aligned}$$

The difference is that $\mathcal{R}(X,Y|Z)$ plays the role of conditional correlation, which is a function of Z; while $\mathcal{R}_Z(X,Y)$ plays the role of partial correlation, which is averaged over Z. But they obey the following inequality:

$$\mathcal{R}_Z(X,Y) \le \sup_z \mathcal{R}(X,Y|Z=z). \tag{5.3}$$

This can be proved by first noticing that for any triple $\{X,Y,Z\}$,

$$\mathrm{Cov}(f(X)-\mathrm{E}[f(X)|Z], h(Y)-\mathrm{E}[h(Y)|Z]) = \mathrm{E}\Big[\mathrm{Cov}(f(X),h(Y)|Z)\Big] \tag{5.4}$$

whenever the needed moments exist. Applying (5.4) to both numerator and denominator of (5.2) (for the two parts in the denominator, we take $f=h$ in (5.4)), we obtain

$$\begin{aligned}\mathcal{R}_Z(X,Y) &= \sup_{f,h\in L^2} \frac{\mathrm{E}\Big[\mathrm{Cov}(f(X),h(Y)|Z)\Big]}{\sqrt{\mathrm{E}\Big[\mathrm{Var}(f(X)|Z)\Big]\mathrm{E}\Big[\mathrm{Var}(h(Y)|Z)\Big]}},\\ &\le \sup_z \mathcal{R}(X,Y|Z=z) \times \sup_{f,h\in L^2} \frac{\mathrm{E}\Big[\sqrt{\mathrm{Var}(f(X)|Z)\mathrm{Var}(h(Y)|Z)}\Big]}{\sqrt{\mathrm{E}[\mathrm{Var}(f(X)|Z)]\mathrm{E}[\mathrm{Var}(h(Y)|Z)]}}\\ &= \sup_z \mathcal{R}(X,Y|Z=z),\end{aligned}$$

where the last equality follows from the Cauchy-Schwartz inequality, which becomes equality when $f=h$.

Now applying (5.3) to (2.6), we have

$$\mathcal{R}_W(\theta, Y_{mis}) \le \sup_w \mathcal{R}(\theta, Y_{mis}|W=w) = r. \tag{5.5}$$

By the Lemma 1 of Yu and Meng (2011),

$$\mathcal{R}(\theta, Y_{mis}) \le \mathcal{R}_W(\theta, Y_{mis}) + (1-\mathcal{R}_W(\theta, Y_{mis}))\mathcal{R}(\theta, W)\mathcal{R}(Y_{mis}, W). \tag{5.6}$$

Noting that under (2.6), $\mathcal{R}(Y_{mis}, W) = \mathcal{R}(\theta, W)$, we see from (5.5)-(5.6) that

$$\mathcal{R}(\theta, Y_{mis}) \le g + (1-g)r, \quad \text{where} \quad g = \mathcal{R}^2(\theta, W). \tag{5.7}$$

Letting $\tilde{Y}_{mis} = Y_{mis} - c\theta$, it is easy to see that the above derivation also applies to $\mathcal{R}(\tilde{Y}_{mis}, \theta)$ and $\mathcal{R}(\tilde{Y}_{mis}, Y_{mis})$, except with r replaced respectively by

$$r_1 \equiv \mathcal{R}(\tilde{Y}_{mis}, \theta|W) = \frac{|c-r|}{\sqrt{1+c^2-2cr}} \quad \text{and} \quad r_2 \equiv \mathcal{R}(\tilde{Y}_{mis}, Y_{mis}|W) = \frac{|1-cr|}{\sqrt{1+c^2-2cr}}, \tag{5.8}$$

where the calculation of $\mathcal{R}(\tilde{Y}_{mis}, \theta|W)$, for example, can be directly read off from the covariance matrix in (2.2) (the missing multiplicative factor W^{-2} is not relevant for the correlation calculation). Consequently, from (1.3), we have

$$\begin{aligned} r_{1\&2} &\le \mathcal{R}(\theta, Y^{mis})\mathcal{R}(\theta, \widetilde{Y}^{mis})\mathcal{R}(\widetilde{Y}^{mis}, Y^{mis}) \\ &\le [g+(1-g)r][g+(1-g)r_1][g+(1-g)r_2] \equiv F(r,c,g), \end{aligned}$$

Now we prove that in the "safe" zone, where $r \le c \le r^{-1}$,

$$F(r,c,g) \le \frac{1}{8}(1+g)^3.$$

We prove this in two steps.

1. For fixed $0 < r, g < 1$, $F(r,c,g)$ is maximized at $c = 1$. Because

$$\frac{\partial F}{\partial c} = \frac{(1-r^2)(1-g)[r+(1-r)g]}{(1+c^2-2cr)^2}[g\sqrt{1+c^2-2cr} + (1-g)(1+c)](1-c), \tag{5.9}$$

we have

$$\frac{\partial F}{\partial c} \begin{cases} > 0, \text{ if } c < 1 \\ = 0, \text{ if } c = 1 \\ < 0, \text{ if } c < 1. \end{cases}$$

Therefore, we see $F(r,c,g)$ is maximized at $c = 1$ for any r and g.

2. For $c = 1$ and fixed g, $F(r,1,g)$ is maximized at $r = \frac{1}{2}$. Because

$$\frac{\partial F(r,1,g)}{\partial r} = \frac{(1-g)[(1-g)\sqrt{(1-r)/2}+g]\left[1-g+\frac{g}{\sqrt{2(1-r)}+1}\right]}{\sqrt{2(1-r)}}(1-2r), \tag{5.10}$$

we see that

$$\frac{\partial F(r,1,g)}{\partial r} \begin{cases} > 0, \text{ if } r < \frac{1}{2} \\ = 0, \text{ if } r = \frac{1}{2} \\ < 0, \text{ if } r < \frac{1}{2}. \end{cases}$$

Hence $F(r,1,g)$ is maximized when $r = \frac{1}{2}$ for any fixed $g < 1$.

As a result, $F(r,c,g) \le F(\frac{1}{2},1,g) = \frac{1}{8}(1+g)^3$.

5.3 Proof of the limits of $\mathcal{R}_v(\theta, W)$

Now we prove that when $W^2 \sim \chi^2_v/v$, $\mathcal{R}_v(\theta, W) \to 1$ as $v \to 0$, under the model (2.6). Consider two functions $g, h : g(\theta) = |\theta|^{v/4}$ and $h(W) = W^{-v/4}$. Because $g(\theta)$ and $h(W)$ both have finite variances, their linear correlation is a lower bound for $\mathcal{R}_v(\theta, W)$. Thus it is sufficient to show that this linear correlation goes to one as $v \to 0$. Direct calculation shows

$$\text{Corr}(|\theta|^{v/4}, W^{-v/4}) \quad = \quad 2^{\frac{v}{4}}\Gamma[\frac{4+v}{8}]\sqrt{\frac{\Gamma[\frac{v}{4}]\Gamma[\frac{v}{2}] - \Gamma[\frac{3v}{8}]^2}{2\pi\Gamma[\frac{v}{2}]^2 - 2^{\frac{v}{2}}\Gamma[\frac{3v}{8}]^2\Gamma[\frac{4+v}{8}]^2}}. \tag{5.11}$$

By the fact that $\Gamma[v]\Gamma[1-v] = \dfrac{\pi}{\sin(\pi v)}$ when $0 < v < 1$, we know $\lim_{v\to 0} v\Gamma[v] = 1$. Together with the fact that $\lim_{v\to 0}\Gamma[\frac{4+v}{8}] = \Gamma[\frac{1}{2}] = \sqrt{\pi}$, we deduce the right-hand side of (5.11) converges to

$$\sqrt{\pi}\sqrt{\frac{8 - \dfrac{64}{9}}{2\pi\times 4 - \dfrac{64}{9}\pi}} = 1,$$

which completes our proof.

The proof for $\mathcal{R}_v(\theta, W) \to 0$ as $v \to \infty$ turns out to be much more involved, even though the result seems obvious because as $v \to \infty$, W converges almost surely to the constant 1, and hence it should be independent of any random variable. The trouble is that there is no theory to automatically guarantee that $\mathcal{R}_v(\theta, W)$ is a continuous function of v. In general, it is a rather complex task to establish even such a continuity with respect to a simple linear combination weight because in general it is not true (see Bryc, Dembo, and Kagan, 2005). We therefore take an indirect route, by considering a two-step Gibbs sampler alternating between sampling $\theta|W$ and $W|\theta$, whose L^2 convergence rate is $\mathcal{R}^2_v(\theta, W)$. It was shown in Roberts and Tweedie (2001) that, for a time reversible Markov chain (such as a two-step Gibbs sampler), its L^2 geometric rate is equivalent to its L^1 rate. By definition, geometric ergodicity in L^1 means that the total variation distance to the target distribution can be bounded by an exponentially decaying function. The bounds in Jones and Hobert (2004) yield precisely such functions, from which we can read off bounds on the geometric rate. Therefore, we can establish the desired result by proving that the L^1 rate converges to zero as $v \to \infty$.

To prove this, we first consider an equivalent two-step Gibbs sampler that alternates between $\theta|Y$ and $Y|\theta$, where $Y = vW^2 \sim \chi^2_v$. Clearly, to draw from $\theta|Y$, we only need to draw $Z \sim N(0, 1)$ independently of Y, and then form $\theta = Z/\sqrt{Y/v}$. To draw $Y|\theta$, we note the identity $Y = (Y + Z^2)/(\theta^2/\nu + 1)$, and the fact that $1/(\theta^2/\nu + 1) = Y/(Y + Z^2)$ has a beta distribution and is independent of $Y + Z^2$, which has a $\chi^2_{\nu+1}$ distribution. Hence we simply draw $G \sim \chi^2_{\nu+1}$ independently of θ and let $Y = G/(\theta^2/\nu + 1)$. Combining the two steps we may represent one iteration of the Y margin by

$$Y \to Y^{new} \equiv \frac{G}{1 + Z^2/Y}, \tag{5.12}$$

where $Z^2 \sim \chi^2_1$, $G \sim \chi^2_{\nu+1}$, and Y, Z^2, G are independent. The Markov chain (5.12) is irreducible (with respect to Lebesgue measure), aperiodic and positive Harris recurrent with χ^2_ν as its invariant

distribution. Therefore, to bound its L^1 rate we can establish suitable minorization and drift conditions and appeal to Rosenthal's (1995) result as stated by Jones and Hobert (2004), Theorem 3.1.

Assume $\nu > 6$ and define

$$V(y) = \nu \left(\frac{\nu - 6}{y} - 1 \right)^2, \quad y > 0.$$

Direct calculation using moments of the inverse χ^2 distribution yields

$$\mathrm{E}[V(Y^{new})|Y] = \gamma V(Y) + b, \quad \text{where } \gamma = \frac{3}{(\nu-1)(\nu-3)} \text{ and } b = \frac{2\nu^2}{(\nu-1)(\nu-3)}.$$

Let $d_R > 4$ be a constant, and suppose ν is large enough so that $\nu > d_R > 2b/(1-\gamma)$. Define the set $C = \{y > 0 : \ V(y) \leq d_R\}$, which is simply the interval

$$y \in [y_*, y^*], \quad \text{where } y_* = \frac{\nu - 6}{1 + \sqrt{d_R/\nu}} \text{ and } \ y^* = \frac{\nu - 6}{1 - \sqrt{d_R/\nu}}.$$

Let $\epsilon = \sqrt{y_*/y^*}$. For any fixed $y \in C$ the density of Z^2/y is bounded below by ϵ times the density of Z^2/y^*, because

$$\sqrt{\frac{y}{2\pi x}} e^{-yx/2} \geq \sqrt{\frac{y_*}{2\pi x}} e^{-y^* x/2}, \quad x > 0.$$

It follows that, if we denote the distribution of $G/(1 + Z^2/y)$ by $P(y, \cdot)$ (i.e., $P(y, \cdot)$ is the transition kernel of (5.12)), then

$$P(y, \cdot) \geq \epsilon P(y^*, \cdot), \quad y \in C.$$

Specifically, one can sample from $P(y, \cdot)$ by setting $Y^{new} = G/(1 + Z^2/y^*)$ with probability ϵ and using another transition rule with probability $1 - \epsilon$.

We have now verified all conditions of Theorem 3.1 of Jones and Hobert (2004) and can conclude that the L^1 rate of (5.12) is bounded above by $\max\{(1-\epsilon)^r, \ U^r/\alpha^{1-r}\}$, where

$$\alpha = \frac{1 + d_R}{1 + 2b + \gamma d_R}, \quad U = 1 + 2(\gamma d_R + b)$$

and $r \in (0, 1)$ is an arbitrary constant. However, for fixed d_R, as $\nu \to \infty$ we have $b \to 2, \ \gamma \to 0, \ \epsilon \to 1$, and this upper bound tends to $5^r/((1+d_R)/5)^{1-r}$. By choosing an arbitrarily large d_R we can make this limiting upper bound arbitrarily small. Hence the L^1 rate must tend to zero as $\nu \to \infty$.

Acknowledgements

This paper is based on a keynote address (by Meng) at the ICMS Workshop on *Advances in Markov Chain Monte Carlo* held in Edinburgh during April 23-25, 2012. We thank the organizers, Mark Girolami, Antonietta Mira and Christian Robert for the invitation and for helpful discussions, many participants—especially Jim Hobert, Omiros Papaspiliopoulos, Gareth Roberts, and David van Dyk—for stimulating exchanges, and the National Science Foundation for partial financial support. We also thank Steven Finch for very helpful proofreading and comments, and Stefam Wilhelm for his timely help regarding the *R*-package *tmvtnorm*.

References

[1] Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics* **17**, 251-269.

[2] Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**, 669-679

[3] Brooks, S., Gelman, A., Jones, G. L. and Meng, X.-L. (Eds) (2011) *Handbook of Markov Chain Monte Carlo.* Chapman & Hall/CRC Press, Boca Raton.

[4] Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *Journal of the American Statistical Association* **80**, 580–619.

[5] Bryc, W., Dembo, A. and Kagan, A. (2005) On the maximum correlation coefficient. *Theory of Probability and its Application* **49**, 132-138.

[6] Gelfand, A. E., Sahu, S. K. and Carlin, B. P. (1995). Efficient parameterisations for normal linear mixed models. *Biometrika* **82**, 479–488.

[7] Gelfand, A. E., Sahu, S. K. and Carlin, B. P. (1996). Efficient parametrizations for generalized linear mixed models. In *Bayesian Statistics* **5**, (Eds: J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith) Oxford University Press, 165–180.

[8] Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409.

[9] Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

[10] Hobert, J. P. (2001a). Discussion of paper by van Dyk and Meng. *Journal of Computational and Graphical Statistics* **10**, 59–68.

[11] Hobert, J. P. (2001b). Stability relationships among the Gibbs sampler and its subchains. *Journal of Computational and Graphical Statistics* **10**, 185-205.

[12] Hobert, J. P. and Marchev, D. (2008). A theoretical comparison of the data augmentation, marginal augmentation and PX-DA algorithms. *Annals of Statistics* **36**, 532–554.

[13] Hobert, J. P. and Roman, J. C. (2011). Discussion of paper by Yu and Meng. *Journal of Computational and Graphical Statistics* **20**, 571-580.

[14] Jones, G. L. and Hobert, J. P. (2004). Sufficient burn-in for Gibbs samplers for a hierarchical random effects model, *Annals of Statistics* **32**, 784-817.

[15] Lange, K and Sinsheimer, J. S. (1993) Normal/independent distributions and their applications in robust regression. *Journal of Computational and Graphical Statistics* **2**, 175-198.

[16] Liu, J. S. and Wu, Y. N. (1999). Parameter expansion for data augmentation. *Journal of the American Statistical Association* **94** , 1264–1274.

[17] Meng, X.-L. and Schilling, S. (1996) Fitting full-information item factor models and an empirical investigation of bridge sampling. *Journal of the American Statistical Association* **91,** 1254-1267.

[18] Meng, X.-L. and van Dyk, D. A. (1997). The EM algorithm — an old folk song sung to a fast new tune (with discussion). *Journal of the Royal Statistical Society B* **59**, 511–567.

[19] Meng, X.-L. and van Dyk, D.A. (1998). Fast EM-type implementations for mixed effects models. *Journal of the Royal Statistical Society B* **60**, 559-578.

[20] Meng, X.-L. and van Dyk, D. A. (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika* **86**, 301–320.

[21] Meng, X.-L. and Xie, X. (2013) I got more data, my model is more refined, but my estimator is getting worse! Am I just dumb? *Econometric Reviews*. To appear.

[22] Papaspiliopoulos, O., Roberts, G. O. and Sköld, M. (2003). Non-centered parameterizations for hierarchical models and data augmentation (with discussion). In *Bayesian Statistics 7* (Eds: J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West), 307-326. Oxford Univ. Press, New York.

[23] Papaspiliopoulos, O., Roberts, G. O. and Sköld, M. (2007). A general framework for the parametrization of hierarchical models. *Statistical Science* **22**, 59–73.

[24] Pena, D., Tiao, G. C. and Tsay, R. S. (eds) (2001) *A Course in Time Series Analysis.* Wiley, New York.

[25] Roberts, G. O. and Tweedie, R. L. (2001). Geometric L^2 and L^1 convergence are equivalent for reversible Markov chains. *Journal of Applied Probability* **38A** (Probability, Statistics and Seismology), 37-41.

[26] Rosenthal, J. S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of the American Statistical Association* **90**, 558-566.

[27] Rosenthal, J. S. (2011). Optimal proposal distributions and adpative MCMC. In *Handbook of Markov Chain Monte Carlo* (Eds: S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng), 93-110. Chapman & Hall/CRC, Boca Raton.

[28] Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82**, 528–540.

[29] Tanner, M. A. and Wong, W. H. (2010). From EM to data augmentation: The emergence of MCMC Bayesian computation in the 1980s. *Statistical Science.* **25**, 506–516.

[30] van Dyk, D. A. and Meng, X.-L. (2001). The art of data augmentation (with discussion). *Journal of Computational and Graphical Statistics* **10**, 1–111.

[31] van Dyk, D. A. and Meng, X.-L. (2010). Cross-fertilizing strategies for better EM mountain climbing and DA field exploration: A graphical guide book. *Statistical Science* **25**, 429–449.

[32] Yu, Y. and Meng, X.-L. (2011). To center or not to center, that is not the question: An ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC efficiency (with discussion), *Journal of Computational and Graphical Statistics* **20**, 531–615.